Sequence diversity analysis of dihydroflavonol 4-reductase intron 1 in common bean

Phillip. E. McClean, Rian K. Lee, and Phillip N. Miklas

Abstract: Variation in common bean (*Phaseolus vulgaris* L.) was investigated by sequencing intron 1 of the dihydroflavonol 4-reductase (*DFR*) gene for 92 genotypes that represent both landraces and cultivars. We were also interested in determining if introns provide sufficient variation for genetic diversity studies and if the sequence data could be used to develop allele-specific primers that could differentiate genotypes using a standard PCR assay. Sixty-nine polymorphic sites were observed. Nucleotide variation (π/bp) was 0.0481, a value higher than that reported for introns from other plant species. Tests for significant deviation from the mutation drift model were positive for the population as a whole, the cultivar and landrace subsets, and the Middle American landrace set. Significant linkage disequilibrium extended about 300 nucleotides. Twenty haplotypes were detected among the cultivated genotypes. Seven recombination events were detected for the whole population, and six events for the landraces. Recombination was not observed among the landraces within either the Middle American or Andean gene pools. Evidence for hybridization between the two gene pools was discovered. Five allele-specific primers were developed that could distinguish 56 additional genotypes. The allele-specific primers were used to map duplicate DFR genes on linkage group B8.

Key words: Phaseolus vulgaris, intron, diversity, evolution.

Résumé : Les auteurs ont mesuré la variation génétique chez le haricot (*Phaseolus vulgaris* L.) en séquençant l'intron 1 du gène codant pour la dihydroflavonol 4-réductase (*DFR*) chez 92 génotypes, une collection représentative tant des variétés de pays que des cultivars. Les auteurs souhaitaient également déterminer si les introns fournissent suffisamment de variation pour réaliser des études de diversité génétique et si ces données rendaient possible le développement d'amorces allèle-spécifiques permettant de distinguer les génotypes par un test PCR. Soixante-neuf sites polymorphes ont été observés. La variation nucléotidique (π/pb) s'élevait à 0,0481, une valeur plus élevée que les valeurs rapportées antérieurement chez d'autres espèces végétales. Une déviation significative par rapport au modèle de dérive génétique a été observée pour la population dans son ensemble, pour les sous-ensembles formés de cultivars ou de variétés de pays, de même que pour le sous-ensemble de variétés d'Amérique centrale. Un déséquilibre de liaison significatif a été noté sur une région de 300 nucléotides. Vingt haplotypes ont été détectés parmi les génotypes cultivés. Sept événements de recombinaison ont été identifiés au sein de l'ensemble de la population et six événements au sein des variétés de pays. Aucun événement de recombinaison n'a été observé au sein des variétés de pays provenant d'Amérique centrale ou des Andes. Des évidences d'hybridation entre ces deux groupes ont été trouvées. Cinq amorces allèle-spécifiques ont été développées permettant de distinguer 56 génotypes additionnels. Ces amorces allèle-spécifiques ont été utilisées pour cartographier des gènes DFR dupliqués sur le groupe de liaison B8.

Mots clés : *Phaseolus vulgaris*, intron, diversité, évolution.

[Traduit par la Rédaction]

Introduction

Sequence diversity is required for any modern molecular analysis in which genotypes must be clearly distinguished. A number of genetic markers systems rely on this underlying diversity. Restriction fragment length polymorphisms are manifested because of sequence differences recognized by a specific restriction enzyme. Randomly amplified polymorphic DNAs (RAPDs) result from the differential genomic distribution of inverted repeats of a unique sequence. Microsatellite (SSR) polymorphisms occur when the copy number of short di- or trinucleotide repeat sequences varies between two genotypes.

The reduction in cost and ease of direct DNA sequencing of target genes has opened up a number of other approaches involving the direct comparison of gene sequences. Direct sequencing uncovers what Rokas and Holland (2000) call RGCs, or rare genomic changes. These define genotypic dif-

Received 1 May 2003. Accepted 26 September 2003. Published on the NRC Research Press Web site at http://genome.nrc.ca on 5 March 2004.

Corresponding Editor: F. Belzile.

P.E. McClean,[1] and R.K. Lee. Department of Plant Sciences, North Dakota State University, Fargo, ND 58105, U.S.A.
P.N. Miklas. USDA–ARS, Vegetable and Forage Crop Research Unit, 24106 North Bunn Rd., Prosser, WA 99350, U.S.A.

[1]Corresponding author (phillip.mcclean@ndsu.nodak.edu)

 doi: 10.1139/G03-103

ferences based on "large-scale mutational changes". These mutations can then be applied to diversity studies at different phylogenetic levels.

Among RGCs, intron-based mutations have a unique utility. Introns are thought to be under less selection pressure than coding regions and consequently accumulate more sequence changes. A single intron may contain multiple insertion–deletions (indels) and single-nucleotide polymorphisms (SNPs). Each of these can be considered a unique marker for a specific allele. In this regard, a single intron can be analogous to a SSR marker that also distinguishes multiple alleles at a given locus. These intron-based alleles can then be surveyed using allele-specific PCR techniques.

Introns generally are easier to work with than other RGCs, such as the mobile SINE and LINE elements, because they are associated with expressed genes that usually are present in a single copy or as a small multigene family. Intron polymorphisms can also be used to map the genetic location of the associated gene. Mobile elements are in higher copy number and more difficult to map. Introns should also distinguish closely related clades within a species. This is not always possible with coding regions because selection pressure has maintained sequence identity. Therefore, intron variability could be useful to consider population genetic parameters relative to mutation, selection, recombination, and linkage disequilibrium.

Common bean (*Phaseolus vulgaris* L.) is a good crop model to test the utility of intron sequence data as RGCs. Two major gene pools were defined for the species based on phaseolin seed storage protein variation (Gepts and Bliss 1986; Gepts 1990), marker diversity (Becerra-Velasquez and Gepts 1994; Koenig and Gepts 1989; Tohme et al. 1996), and morphology (Gepts and Debouck 1991). The Middle American gene pool extends from Mexico through Central America and into Colombia and Venezuela, whereas the Andean gene pool is found in southern Peru, Chile, Bolivia, and Argentina. The two domesticated gene pools appear to converge in Colombia (Gepts and Bliss 1986). Races have also been defined for each gene pool (Singh et al. 1991). The Middle American gene pool consists of races Durango, Jalisco, and Mesoamerica, whereas races Nueva Granada, Chile, and Peru represent the Andean gene pool. Based on a novel phaseolin type, a third, possibly ancestral, wild gene pool based in southern Ecuador and northern Peru was described (Debouck et al. 1993; Kami et al. 1995).

The primary goal of this research is to investigate the utility of intron sequence data for both phylogenetic and population genetic studies by analyzing both closely and distantly related common bean clades. We studied intron 1 of the gene coding for the enzyme dihydroflavonol 4-reductase (DFR) as our model. The enzyme is necessary for the anthocyanin biosynthetic pathway that produces the many colors and hues found in the seed coats of common bean. A large collection of landraces and cultivars representing all the races of the species was analyzed. This allowed us to uncover a number of indels and SNPs useful for phylogenetic and population genetic analyses. The sequence diversity data were also used to develop allele-specific primers that were used to assess marker diversity among another set of genotypes. The primers were also used to discover the map position of duplicated DFR genes.

Materials and methods

Plant material

A collection of 92 common bean genotypes, consisting of 58 cultivars and breeding lines (Table 1) and 34 landraces (Table 2), were analyzed. The wild accession DGD-1962 was included because it represents a genotype that may have contributed to the evolution of common bean (Kami et al. 1995). The cultivars were selected because they represent heirloom or historical genotypes, ancestors to many of the modern dry bean cultivars grown in the U.S.A. (McClean et al. 1993), elite genotypes currently grown in the U.S.A., the current snap bean germplasm, donors of important agronomic genes, or parents of the most important mapping populations (Kelly and Miklas 1999). The landraces are primarily members of a USDA National Plant Germplasm System collection that represents diversity across the six common bean races (http://www.ars-grin.gov/cgi-bin/npgs/html/dno_eval_acc.pl?83036+490969+30). Additional landraces were included because they are of historical importance or are important trait donors.

DNA isolation, amplification, cloning, and sequencing

DNA was extracted from young, expanding leaf tissue using the method of Doyle and Doyle (1990). The dihydroflavonol 4-reductase nucleotide sequences from *Glycine max* (AF167556), *Medicago truncatula* (AW981263), *Arabidopsis thaliana* (At5g42800), and *Fragaria* × *ananassa* (AF029685) were aligned. Based on conserved nucleotide and derived amino acid sequence variability, the region spanning intron 1 was targeted for amplification. The PCR primers were based on the following amino acid sequences: N-terminal, RATV/LRD; C-terminal, GVFHVA. The redundant primer sequences only represented the nucleotide diversity in the reference sequences. The sequence of those primers are: forward, 5′-CGWGCCACCGTDCKMGA-3′; reverse, 5′-GCVAMRTGRAAMACWCC-3′. Fifteen nanograms of DNA from genotype 5-593, a Florida breeding line, was initially amplified using the reagent concentrations described in Brady et al. (1998) and the following amplification conditions: 94 °C for 3 min; 45 cycles of 94 °C for 1 min, 57 °C for 1 min, and 72 °C for 2 min; 1 cycle of 72 °C for 5 min. The amplified fragment was cloned using the pGEM-T Easy Vector System 1 (Promega Corporation, Madison, Wis.). The cloned fragment was sequenced in both directions using a Beckman CEQ 2000XL DNA Analysis System following dideoxy-chain terminator sequencing with the Beckman CEQ DTCS sequencing kit (Beckman Coulter Inc., Fullerton, Calif.). The derived nucleotide sequence was used to develop the following *Phaseolus*-specific DFR primers: forward (Pv-DFRi1-F), 5′-CGTGCCACCGTTCTC GACCCAG-3′; reverse (Pv-DFRi1-R), 5′-AATGCTTCACCTTCTTCATGTTCC-3′. These primers were subsequently used to amplify DNA fragments from all other genotypes using the same PCR amplification conditions described for the redundant primers. The PCR fragments were sequenced in both directions in the same manner as the clones, except the *Phaseolus*-specific DFR primers were used. The DNA sequence chromatograms were analyzed using the Staden Package (Staden 1994; http://www.mrc-lmb.cam.ac.uk/pubseq/staden_home.html). If multiple heterozygous nucleo-

Table 1. *Phaseolus vulgaris* cultivars and breeding lines used in this study.

Genotype	Accession No.[a]	Market class	Gene pool[b]	DFR haplotype
'Dorado'		Small red	Middle American	1
'Tio Canela'		Small red	Middle American	1
'Aztec'	PI561473	Pinto	Middle American	2
'Cantare'		Snap	Middle American	2
'Fiesta'	PI550129	Pinto	Middle American	2
'Montrose'	PI612595	Pinto	Middle American	2
'Othello'	PI578268	Pinto	Middle American	2
'Topaz'		Pinto	Middle American	2
'UI 114'		Pinto	Middle American	2
'Montcalm A'		Kidney	Andean	3
5-593		Black	Middle American	5
'Coulee'		Medium red	Middle American	5
'Emerson'	PI615391	Great northern	Middle American	5
'ICA-Bunsi'		Navy	Middle American	5
'Seafarer'	PI549871	Navy	Middle American	5
'Sutter Pink'		Pink	Middle American	5
'NEP-2'		Navy	Middle American	6
'Aurora'	PI289445	Navy	Middle American	7
'C-20'	PI550133	Navy	Middle American	7
'Domino'	PI550273	Black	Middle American	7
'Mayflower'	PI531235	Navy	Middle American	7
'A 55'	PI632407	Black	Middle American	8
'BAC6'		Mottled	Middle American	8
'BAT93'		Light tan	Middle American	8
'Blue Lake Stringless' FM1 (BLS FM1)		Snap	Middle American	8
'HT7719'		Black	Middle American	8
'Kentucky Wonder 780'		Snap	Middle American	8
ND-88-106-04		Navy	Middle American	8
'Newport'	PI586656	Navy	Middle American	8
'NW63'	PI550016	Medium red	Middle American	8
'Raven'	PI578078	Black	Middle American	8
'UI 37'		Medium red	Middle American	8
'Vista'	PI559389	Navy	Middle American	8
'Viva'	PI549940	Pink	Middle American	8
'Arthur'		Navy	Middle American	9
'CDRK 82'		Kidney	Andean	10
'Redkloud'	PI599028	Kidney	Andean	10
'Olathe'	PI550027	Pinto	Middle American	11
'Sprite'	PI550248	Snap	Andean	11
'Cardinal'		Cranberry	Andean	16
'Enola'	PI596310	Yellow	Andean	16
'ICA-Viboral'	G12722	Cranberry	Andean	16
'Montcalm B'		Kidney	Andean	16
'Calima'	G1853, PI310511	Red mottled	Andean	17
'91G'		Snap	Andean	20
'Benton'		Snap	Andean	20
'Bill Z'	PI522246	Pinto	Middle American	20
'Contender'	PI52779	Snap	Andean	20
'Harvester'	PI549648d	Snap	Andean	20
NY-6020-4		Snap	Andean	20
'Taylor Horticultural'		Cranberry	Andean	20
'Tender Crop'	PI549632	Snap	Andean	20
'Wagenaar'		Yellow	Andean	20

[a]The Centro Internacional de Agricultura Tropic (CIAT) accession numbers are preceded by the letter G. The USDA National Plant Germplasm System (NPGS) accession numbers are preceded by the letters PI.

[b]Gene pool designation is based on seed size, color, and pattern and follows the convention of Singh et al. (1991).

Table 2. *Phaseolus vulgaris* landraces used in this study.

Genotype	Accession No.[a]	Race[b]	Collection Country, State[c]	DFR haplotype
Mexico 309	G3825	Mesoamerica	Mexico, Mexico	1
Orgulloso	G14027, PI608378	Mesoamerica	Nicaragua	1
Bayo	PI313540	Durango	Mexico, Federal District	2
Cejita	G1796, PI309774	Jalisco	Mexico, Jalisco	2
Durango 222	G18440, PI608380	Durango	Mexico, Durango	2
Flor de Mayo	PI309707	Jalisco	Mexico, Mexico	2
Flor de Mayo IV	G22036, PI608387	Jalisco	Mexico	2
Frijola	G22039, PI608385	Jalisco	Mexico	2
Garbancillo Zarco	G15821, PI608386	Jalisco	Mexico, Jalisco	2
Guanajuato 31	G2618, PI608383	Durango	Mexico, Guanajuato	2
Ojo de Cabra Santa Rita	G22078, PI608382	Durango	Mexico	2
Zacatecano	G2858, PI608381	Durango	Mexico, Jalisco	2
Jamapa	G1459, PI268110	Mesoamerica	Mexico, Veracruz	4
Anitoquia 106	PI313580	Nueva Granada	Colombia	8
Brasil 2	G3807, PI608377	Mesoamerica	Brasil	8
Carioca	G4017, PI608375	Mesoamerica	Brasil	8
Dilmason	G577, PI608384	Durango	Turkey, Konya	8
Ecuador 299	G2571, PI313691	Mesoamerica	Ecuador, Morona Santiago	8
Porrillo Sintético	G4495, PI608376	Mesoamerica	El Salvador, La Paz	8
Black Turtle Soup (T39)	G17640	Mesoamerica	U.S.A.	9
Rio Tibagi	G4830, PI608379	Mesoamerica	Brazil, Santa Catarina	9
Bolon Bayo	G12230, PI608404	Peru	Ecuador, Chimborazo	12
Coscorron Corriente	G50622, PI608396	Chile	Chile	13
Don Timoteo	G13936	Chile	Chile	13
Frutilla Corriente	PI608395	Chile	Chile	13
Tórtolas Corriente	G24554, PI608397	Chile	Chile	14
Mortiño	G4553, PI608400	Peru	Colombia, Antioquia	15
Bolón Rojo	G12209, PI608403	Peru	Ecuador, Chimborazo	16
Nuña Mani Roja	G12582, PI608402	Peru	Peru, Cajamaraca	16
Coscorrón	PI557465	Chile	Chile	18
Nuñas	PI531862	Peru	Peru, Miraflores	18
Caballero	PI608401	Peru	Peru	19
Ecuador 1056	PI608391	Nueva Granada	Ecuador	19
Jalo EEP 558	G9603, PI608392	Nueva Granada	Brazil, Minas Gerais	19
Alubia Cerrillos	G7930, PI608394	Nueva Granada	Peru, La Libertad	20
Blanco Español	PI608398	Chile	Chile	20
Jatu Rong	G122, PI608390	Nueva Granada	India, Punjab	20
Pompadour Checa	G18264, PI603944	Nueva Granada	Dominican Republic	20
Radical San Gil	G24536, PI608393	Nueva Granada	Colombia, Santader Del Sur	20
DGD-1962 (wild)	G21245		Peru, Cajamarca	21

[a]The Centro Internacional de Agricultura Tropic (CIAT) accession numbers are preceded by the letter G. The USDA National Plant Germplasm System (NPGS) accession numbers are preceded by the letters PI.
[b]Race designation follows the convention of Singh et al. 1991.
[c]Collection information was obtained from either the CIAT (http://webpc.ciat.cgiar.org:8080/urg/beans.htm) or the USDA NPGS (http://www.ars-grin.gov/npgs/acc/acc_queries.html) databases.

tide sites were observed, or if the sequence reads terminated early, the amplification fragement was cloned and multiple clones were sequenced. This was only necessary for the landrace 'Jalo EEP 558' and the red kidney bean cultivar 'Montcalm'.

Population genetics and phylogenetic tree analysis

All sequences (GenBank accession Nos. AY348596-AY348688) were initially aligned using MultAlin software (Corpet 1988) available from http:// prodes.toulouse.inra.fr/multalin/multalin.html. Individual sequence chromatograms were reexamined using the Staden package to ensure accuracy of low-frequency (three or less) polymorphic nucleotides. If the sequence was unclear, the fragment was resequenced. Following corrections, the sequences were realigned using ClustalX Ver. 1.81 (Thompson et al. 1997). The DnaSP (Rozas and Rozas 1999) software package was used to calculate population genetic parameters. Nucleotide diversity (π) was estimated as the mean pairwise nucleotide differences (Nei 1987), and nucleotide polymorphism (θ) was estimated as the expected number of polymorphic sites per sequence (Watterson 1975). To test if the sequences deviated from the mutation-drift model, the Fu and Li (1993) D^* and F^* parameters and the Tajima (1989) D parameter

were estimated. Indel sites were excluded for the purpose of these calculations. The minimum number of recombination events was estimated using the four-gamete test (Hudson and Kaplan 1985). Linkage disequilibrium was estimated by performing the Fisher's Exact Test for all pairwise comparisons of polymorphic sites. To account for multiple comparisons, the Bonferroni (see Weir 1996) correction was applied.

Neighbor-joining (NJ) and maximum parsimony (MP) trees were developed using PAUP (version 4.0b10). For NJ, the HKY85 distance model for nucleotide substitutions was used. A bootstrap analysis of the tree was performed with 1000 resamplings. For the MP analysis, each substitution was equally weighted, and indels were considered a character. One thousand heuristic searches with stepwise random addition of sequences were employed to discover the equally parsimonious trees. A 50% majority rule consensus tree was derived from the equally parsimonious trees.

Allele-specific primers and linkage mapping of *DFR*

The sequence information from this collection of genotypes was used to design additional primers to amplify allele-specific DFR fragments. Those primers are as follows: Andean-DFR-F, 5′-GGTCTGGTGGATCTTTGTTGGTGC-3′; Durango-DFR-F, 5′-GTATTATCATGTAGGGTCTGATGG-3′; Dorado-DFR-R, 5′-GGTACTTGGGTACTCCAATCCTT-3′; Jamapa-DFR-F, 5′-TGTTTTTGGTTTTTGTATAAGGATTTG-3′; Aurora-DFR-F, 5′-TGTTTTTGGTTTTTGTATAAGGATTTA-3′; and DFRi1-internal-R, 5′-CAGTGATAACATGAAAGTGTTAGGTTG-3′. Combinations of these primers were used to amplify specific *DFR* intron 1 alleles using the basic amplification conditions described above. The primer combinations, DNA annealing temperatures, amplification cycles, expected product sizes, and the haplotypes that amplify the product are shown in Table 3.

The genetic location of *DFR* loci was determined by screening a recombinant inbred population (n = 79) developed by crossing 'Dorado' (tested as DOR 364) and 'Xan 176'. 'Dorado' is homozygous for the *Dorado* allele, whereas 'Xan 176' is homozygous for the *Durango*-331 allele. Individual members of the population were scored for both the *Durango* and *DoradoI*-331 alleles. Linkage mapping was performed with MAPMAKER 3.0 (Lander et al. 1987). Markers were assigned to a linkage group using the "group" command with distance set to less than 30 cM and a minimum LOD score of 3.0. Markers within a linkage group were ordered using the "compare" and "try" command, and the order was verified using the "ripple" command with a LOD value of 3.0.

Results

Cloning *DFR* intron 1 sequences from common bean

Amplification of DNA of common bean genotype 5-593 produced three fragments that were approximately 125, 300, and 700 bp in size. All three fragments were cloned and sequenced. Because only the largest fragment, 689 nucleotides in length, was homologous to other DFR genes, only it was studied further. The sequences of eight different clones were identical. A GenBank BLASTx search using the sequence of this fragment as the query determined that the amino acid sequence corresponding to nucleotides 556 to 675 were most homologous to 40 amino acids of the DFR protein from *G. max* (AAD54273; $E = 6 \times 10^{-15}$) and *M. sativa* (P51109; $E = 6 \times 10^{-13}$). Within this region, the common bean fragment was 95% and 85% identical to *G. max* and *M. sativa*, respectively. A BLASTx search of the MIPS *Arabidopsis* database (http://mips.gsf.de/proj/thal/db/search/search_frame.html) revealed the common bean sequence was homologous to the *A. thaliana* DFR protein (At5g42800; $E = 5 \times 10^{-12}$; 72.5% identical in the coding region). Further analysis revealed these sequences corresponded to the first 40 amino acids fully encoded by exon two of the *A. thaliana* gene. These results convinced us that we had successfully cloned a common bean fragment containing intron 1 and part of exon 2 of *DFR*.

In addition to *A. thaliana*, *DFR* intron 1 and (or) exon 2 sequences are available for *Fragaria vesca* subsp. *vesca* (AY017480), *Fragaria nubicola* (AY017487), *Oryza sativa* subsp. *japonica* (AB003495), *Triticum aestivum* (AY210885), and a *Lophopyrum ponticum* × *Triticum aestivum* hybrid (AY208999). Common bean and all of these sequences shared one common feature: intron 1 was located between the second and third nucleotide of the same amino acid of the protein sequence; however,the amino acid at this site varied. For these other species, the intron ranged in size from 76–115 nucleotides. By comparison, the 5-593 intron sequence was 531 nucleotides. To determine if the longer-sized intron was unique to common bean or similar in related legumes, the redundant primers were used to amplify corresponding fragments from soybean (*G. max* 'Barnes') and cowpea (cultivar unknown). Sequences of cloned amplified fragments determined that soybean contained multiple, unique *DFR* fragments with intron 1 lengths of 138, 159, and 452 nucleotides. By contrast, only a single sequence was detected in cowpea, and the intron length was 149 nucleotides. We attempted to align the intron sequences from all of the species, but no discernable homology was detected.

DFR intron 1 diversity in common bean

The 92 genotypes in this study are a collection of cultivars and landraces that represent much of the diversity within the cultivated form of *P. vulgaris*. All of these genotypes were monomorphic within the coding region of the amplified DFR fragment. By contrast, 69 of the 552 comparative intron 1 sites were polymorphic (Table 4). Each site was dimorphic except for nucleotide 407. Only two singleton sites were discovered. In addition, 28 indels were observed. Although a number of polymorphic sites and indels (nucleotides 135–142, 315, and 373) distinguished the Middle American and Andean genotypes, other polymorphisms were shared across the two major gene pools. For example, one indel (nt 312–314) discriminated 'Durango' and 'Jalisco' cultivars and landraces from the other Middle American race (Mesoamerica) and the three Andean races.

Nucleotide variation among all genotypes was estimated as diversity (π/bp; Tajima 1989) and polymorphism (θ/bp; Watterson 1975). These values were 0.0481 and 0.0256, respectively. Tajima's D (Tajima 1989) and Fu and Li's D^* and F (Fu and Li 1993) tests were conducted to determine if the polymorphisms deviated from the mutation-drift model (Kimura 1983). Significant positive values were observed for each statistic indicating an excess of intermediate frequency polymorphism that can be the result of either balancing se-

Table 3. Allele-specific primers, amplification conditions, product size, and haplotype scores for *P. vulgaris* dihydroflavonol 4-reductase intron 1.

Allele	Primers	T_A (°C)	Cycles	Product size	Haplotype(s) with product
Andean	Forward: Andean-DFR-F Reverse: DFRi1-internal-R	55	45	322	10–20
Aurora	Forward: Aurora-DFR-F Reverse: DFRi1-internal-R	63	35	293	2, 6, 7
Dorado	Forward: Pv-DFRi1-F Reverse: Dorado-DFR-R	56	25	391	1
Durango	Forward: Durango-DFR-F Reverse: DFRi1-internal-R	55	25	160 331	1, 2 8
Jamapa	Forward: Jamapa-DFR-F Reverse: DFRi1-internal-R	51	25	293	1, 3–5, 8–20

lection, diversifying selection or population subdivision (Hartl and Clark 1997). To consider these possibilities, the data was subdivided relative to breeding history (cultivar or landrace) and (or) gene pool (Middle America or Andean), and the population genetics statistics were recalculated (Table 4).

Nucleotide diversity values for both cultivars and landraces were equivalent to that observed for the entire population. This pattern was also observed when the genotypes were analyzed within gene pools. The Middle American gene pool was much more diverse than the Andean. Between the two Middle American forms, only the landraces deviated from the mutation-drift model. In contrast, the Andean population as a whole and each of the two forms fit the model.

Haplotype variation

Collectively, the sequences of the 92 genotypes defined 20 haplotypes (Fig. 1). In addition, the wild *P. vulgaris* genotype DGD-1962 represented an additional haplotype. The relationships among the haplotypes are represented by MP and NJ trees (Fig. 2). The haplotypes, (displayed in Fig. 1 in the same order (except haplotype 21) as they are in the two trees), fall into three major groups: 1, 2, and 21; 3–9; and 10–20.

The two trees (Fig. 2) show that the later two haplotype groups form tight clusters with relatively little diversity. All Mesoamerican genotypes are a member of haplotypes 3–10. By contrast, the highly related haplotypes 10–20 represent the three Andean races (Chile, Peru, and Nueva Granada). Among the Mesoamerican race, haplotypes 4 and 8 were the most diverse, yet only differed by four single nucleotide polymorphisms (SNPs) and one indel event. Among the Andean genotypes, haplotype 13 has the largest collection of race Chile genotypes, whereas haplotypes 16 and 20 contain the largest collection of race Peru and Nueva Granada genotypes, respectively. Among these, the largest diversity was noted between haplotypes 13 and 16 (three SNPs and one indel).

Haplotype 1 primarily consists of the small red beans of race Mesoamerica. Haplotype 2 consists of only 'Durango' and 'Jalisco' genotypes. These two haplotypes are quite distinctive and differ by 16 SNPs and 4 indel events. Interestingly, the race Mesoamerican haplotypes (4–9) clusters with the Andean haplotype cluster before these two clusters are grouped with either Middle American haplotypes 1 or 2 (Fig. 2).

The wild genotype DGD-1962 was included because it is suggested to represent an ancestral common bean based on its phaseolin gene structure (Kami et al. 1995). In this analysis, it shares the major deletion event (nt 135–142) with the Middle American genotypes (Fig. 1). Among the representative haplotypes, DGD-1962 is most similar to the Durango/Jalisco haplotype 2 (15 SNPs and two indel differences) and is most distant from Andean haplotype 13 (45 SNPs and five indels).

Recombination and linkage disequilibrium within *DFR* intron 1 of common bean

Recombination also affects haplotype diversity. The minimum number of recombination events (R_M) was calculated using the four-gamete test (Hudson and Kaplan 1985; Table 4). When the entire population was analyzed, $R_M = 7$. This was reduced to $R_M = 6$ when only the landraces were analyzed. When the genotypes within a gene pool were considered, this value was reduced further. Finally, among the landraces within a gene pool, recombination was not observed.

Two methods were used to measure linkage disequilibrium. First the linkage disequilibrium statistic D (Lewontin and Kojima 1960) was calculated for all informative sites. Fisher's exact test was performed to determine the significance of each comparison. Despite the large number of recombination events, there were still a large percentage of significant comparisons. Over half of these comparisons were significant after the Bonferroni correction was applied. These same linkage disequilibrium trends persisted when the cultivar or landrace genotypes were considered as a test group. When the genotypes were analyzed within gene pools, a large percentage of comparisons were significant for all the Middle American genotypes. This was also observed for the Middle American cultivars. Finally, the Z_nS statistic (Kelly 1997), which uses the average pairwise allelic correlation coefficient, was calculated. This statistic was significant for all Middle American comparisons, but all other comparisons were not significant.

Linkage disequilibrium was also considered with respect to distance. The linkage disequilibrium statistic of Hill and Robertson (1968), R^2, was plotted relative to distance (Fig. 3). This statistic is equivalent to the square of the cor-

Table 4. Population genetic statistics for intron 1 of the dihydroflavonol 4-reductase gene of common bean.

				Gene pool					
				Middle America			Andean		
	All[a]	Cultivars	Landraces	All	Cultivars	Landraces	All	Cultivars	Landraces
No. of sequences	92	53	39	56	35	21	36	18	18
No. of total sites	552	552	552	552	552	552	552	552	552
Sites less indels	524	525	524	527	525	527	539	541	539
Indels	28	27	28	25	25	25	13	11	13
Polymorphic sites	69	67	68	45	45	41	10	6	10
Singletons	2	1	1	4	4	0	2	0	2
Haplotypes	20	13	13	9	8	5	11	5	8
Nucleotide variation									
Diversity (π/bp)	0.0481	0.0463	0.0495	0.0314	0.0270	0.0347	0.0048	0.0039	0.0054
Polymorphism (θ/bp)	0.0259	0.0281	0.0307	0.0186	0.0207	0.0216	0.0045	0.0032	0.0054
Neutrality tests									
Tajima's *D*	2.8124***	2.2477*	2.2214*	2.3352*	1.0844	2.3272*	0.2401	0.7045	–0.0310
Fu and Li's *D**	1.9376**	1.9273**	1.8400**	1.1872	1.1932	1.7099**	0.2719	1.2590	0.4947
Fu and Li's *F**	2.7551**	2.4452**	2.3466**	1.9240*	1.3697	2.2136**	0.3073	1.2732	0.4002
Recombination									
# four gamete type site pairs	193	152	93	23	17	0	9	3	0
R_m[b]	7	7	6	3	2	0	2	1	0
Linkage disequilibrium									
Pairwise comparisons	2211	2145	2211	820	820	820	28	15	28
% Fisher's exact test[c]	76.3	67.4	66.4	86.5	63.9	63.9	25.0	26.7	17.9
% Bonferroni[d]	53.3	37.0	40.7	63.9	47.4	39.1	21.4	20.0	00.0
Z_nS[e]	0.3183	0.3341	0.3411	0.5241**	0.4990**	0.5898**	0.3280	0.3316	0.2363

[a]Significance levels ***, $P < 0.01$, **, $P < 0.02$; *, $P < 0.05$.
[b]Minimum number of recombination events based on the four-gamete test.
[c]Percentage of significant ($0.05 < P$) pairwise comparisons by the Fisher's Exact Test.
[d]Percentage of significant ($0.05 < P$) pairwise comparisons by the Fisher's Exact Test after the Bonferroni correction for multiple comparisons.

Fig. 1. Nucleotide polymorphisms among the *Phaseolus vulgaris* intron 1 of dihydroflavonol 4-reductase. Haplotype 1 was arbitrarily considered the reference allele. Dashes (–) indicate nucleotide identity. Dots indicate an insertion–deletion event. See Tables 1 and 2 for genotypes that are members of each haplotype.

```
H                                  Polymorphic DFR nucleotide
a  ------------------------------------------------------------------------------------------------------------------
p                 1111111111111111111111111111122222222222233333333333333333333333333333333344444444444444444455
1  22245567788888000111122333333444466788888890135667899990011111222233344556667777777880000122225677789 33
o  58927801501356567046704456789012328502567868114290901686812345245826969293560123458142679601367034532 12
------------------------------------------------------------------------------------------------------------------
 1  AATGCGTACCACTAGAGTGTGGTT........AG.GGACCCTCTTCGGTGTAGATGCATGGATAGCGCAGTTGGAGCC....GCATACAA..TATA...CAG.
 2  ---------T----T---C-----........-AAA---T-C-C.---------C--------G----GA--A-----.CTG--C---G-G.---TAA.---.
21  ---------T-A------CC-A--........-A---------C.---------C-T------G-TA-G--------TTCTG--C--TG-G.---TA..---.
 3  CTCA-T---T-----GA-C---CA........G-G-A----.TC.T----G---C---...--G--AT--AG----T-.CTGATCCGAG-A.CG-TAA.AT-.
 4  CTC------T-----GA-C---CA........G-G------.TC.T----G---C---...--G--AT--AG----T-.CTGATCC--GG..-G.T...A-TT
 5  ---------T-----GA-C---CA........G-G------.TC.T----G---C---...--G--AT--AG----T-.CTGATCC--GG..-G.T...A-TT
 6  ---A-T---T-----GA-C---CA........G-G-A----.TC.T----G---C---...--G--AT--AG----T-.CTGATCC--GG..-G.T...A-TT
 7  ---------T-----GA-C---CA........G-G-A----.TCTT----G---C---...--G--AT--AG----T-.CTGATCC--GG..-G.T...A-TT
 8  CTCA-T---T-----GA-C---CA........G-G--G---CTC.T----G--AC---...--G--AT--AG----T-.CTGATCC--GG..-G.T...A-TT
 9  CTCA-T---T-----GA-C---CA........G-G-AG---.TC.T----G--AC---...--G--AT--AG----T-.CTGATCC--GG..-G.T...A-TT
10  CTC----GTT--AG-G--C-A-C-TTGTTGGTGA.A-G---CTC.-ATCAG-AGC--G....CGA-A-G----AGA--..TG--CCGAG-A.CG-TAAA---.
11  CTC---CGTT--AG-G--C-A-C-TTGTTGGTGA.A-G---CTC.-ATCAG-AGC--G....CGA-A-G----AGA--..TG--CCGAG-AACG-TAAAA--.
12  CTC-G-CGTT--AG-G--C-A-C-TTGTTGGTGA.A-G---CTC.-ATCAGGAGC--G....CGA-A-G----AGA--..TG--CCGAG-A.CG-TAAAA--.
13  CTC-G-CGTT--AG-G--C-A-C-TTGTTGGTGA.A-G---CTC.-ATCAGGAGCA-G....CGA-A-G----AGA--..TG--CCGAG-AACG-TAAAA--.
14  CTC-G-CGTT--AG-G--C-A-C-TTGTTGGTGA.A-G--TCTC.-ATCAGGAGCA-G....CGA-A-G----AGA--..TG--CCGAG-AACG-TAAAA--.
15  CTC-G-CGTTG-AG-G-GC-A-C-TTGTTGGTGA.A-G---CTC.-ATCAGGAGC--G....CGA-A-G----AGA--..TG--CCGAG-A.CG-TAAAA-T.
16  CTC-G-CGTTG-AG-G-GC-A-C-TTGTTGGTGA.A-G---CTC.-ATCAGGAGC--G....CGA-A-G----AGA--..TG--CCGAG-A.CG-TAAAA--.
17  CTC---CGTTG-AG-G-GC-A-C-TTGTTGGTGA.A-G---CTC.-ATCAGGAGC--G....CGA-A-G----AGA--..TG--CCGAG-A.CG-TAAAA--.
18  CTC----GTTG-AG-G-GC-A-C-TTGTTGGTGA.A-G---CTC.-ATCAGGAGC--G....CGA-A-G----AGA--..TG--CCGAG-A.CG-TAAA---.
19  CTC-G-CGTT--AG-G-GC-A-C-TTGTTGGTGA.A-----CTC.-ATCAGGAGC--G....CGA-A-G----AGA--..TG--CCGAG-A.CG-TAA.A--.
20  CTC-G-CGTT--AG-G--C-A-C-TTGTTGGTGA.A-G---CTC.-ATCAG-AGC--G....CGA-A-G----AGA--..TG--CCGAG-A.CG-TAAAA--.
```

Fig. 2. Phylogenetic trees inferred from dihydroflavonol 4-reductase intron 1 sequence data. For each, the wild *P. vulgaris* genotype DGD-1962 was used to root each tree. The haplotype designation is shown in parentheses at the end of each node. (A) The 50% majority rule consensus MP tree derived from the 126 equally parsimonious trees. Length = 144 steps; consistency index = 0.757; retention index = 0.930. The numbers at the node represent the percentage of times that a genotype pattern occurs among the 126 trees. If the percentage was less that 50%, the node was collapsed. (See B for the complete list of genotypes that comprise the haplotype.) (B) The NJ distance tree. The numbers represent the percentage bootstrap value (out of 1000 resamplings) for a particular node. Only nodes with percentage bootstrap values greater than 50 are shown. The genotypes at each branch tip have identical sequences and represent a single haplotype. Landraces are listed in italics.

relation coefficient between alleles at two loci. In the present context, each nucleotide is considered to be a locus. From this representation, it appears that significant linkage disequilibrium is maintained over a distance of about 300 nucleotides in intron 1 of DFR. This same pattern was observed when the cultivar and landrace genotypes were analyzed as individual groups. Further analysis revealed the same pattern for the Middle American genotypes, whereas no pattern of linkage disequilibrium over the length of the intron was noted for the Andean genotypes.

Intron-based marker diversity of *DFR*

DNA sequence data provides the necessary information to develop allele-specific primers that can be used to characterize genotypes based on distinctive features such as indels and SNPs. To demonstrate this principle with the *DFR* intron sequence, we searched for major variants that distinguished genotypes of different genetic backgrounds. These primer pairs were applied to a set of 56 genotypes that had not been sequenced and represent an array of common bean market classes (Table 5). The first primer pair targeted the insertion at nt 135–142 (*Andean* allele; Fig. 1) that distinguishes Andean (haplotypes 10–20) and Middle American (haplotypes 1–9, 21) genotypes. All of the coscorron and kidney genotypes, known to be of Andean ancestry, were positive for this allele. In addition, the pinto genotype 'Mestizo' was also positive for this allele. This is unexpected because pintos are generally considered to be of Middle American origin. 'Mestizo' is a hybrid between 'Bayo Victoria' and 'Olathe' (Acosta-Gallegos et al. 2001). Although a pinto cultivar, the pedigree of 'Olathe' contains the Andean cultivar 'Golden Gate Wax' that contributed the *Ur-6* bean rust resistance gene. The pedigree explains the sequence analysis result that shows that 'Olathe' is positive for the *Andean* allele. Therefore, 'Mestizo' probably received its *DFR* allele from 'Olathe'.

Because sequence diversity was greater within the Middle American than the Andean gene pool, it was targeted for additional allele-specific primer development. A primer pair, based on the insertion at nt 311–313 (Fig. 1), was created to amplify the *Durango* allele. That allele was only found in haplotypes 1, 2, and 21. Except for 'Mestizo' and 'Olathe' (see above), all genotypes of the pinto market class were positive for this allele. This is consistent with the sequence data for the pinto genotypes. In addition, about half of the great northern genotypes were also positive for this allele. The pinto cultivar 'Sierra' is a major contributor to the development of the great northern cultivar 'Matterhorn' (Kelly et al. 1999) and probably donated its *DFR* allele to this genotype. The great northern breeding lines G97918 and G97943 mostly likely obtained the 'Durango' allele from 'Matterhorn', one of the parents for each line.

To test the utility of SNP-based allele-specific primers, we targeted polymorphisms at nt 162, 168, 175, and 180 (Fig. 1). The first three nucleotide positions distinguish the Mesoamerican genotypes from all the other races. In theory, the SNP at nt 180 distinguishes Mesoamerican haplotypes 4, 5, and 8 (*Jamapa* allele) from 6, 7, and 9 (*Aurora* allele). We employed the SNAP procedure (Drenkard et al. 2000) and the SNAPER software (http://ausubellab.mgh.harvard.edu) to design appropriate primers for the two alleles. These two forward primers differ at the 3′ end. In addition, a mismatch to the original sequence was added at the next to the last nucleotide.

As found with the sequencing data, a number of genotypes from different market classes contain the *Jamapa* allele. Among the great northern cultivars, 'GN1140' is positive for this allele. This is consistent with the sequencing results for 'Emerson', a cultivar that has GN1140 as a parent. GN1140 is also part of the pedigree of 'Harris' and 'Weihing' (Coyne et al. 2000), two other Great northern cultivars that possess the *Jamapa* allele. 'Alberta Pink', 'UI537', and the breeding line L94C356 are pink-seeded genotypes that also possess the *Jamapa* allele. This is in agreement with the sequencing data for the pink cultivars 'Viva' and 'Sutter Pink'. Finally, as observed with the sequencing data, all of the red cultivars with medium-sized seed (formerly Red Mexican) also possess the *Jamapa* allele.

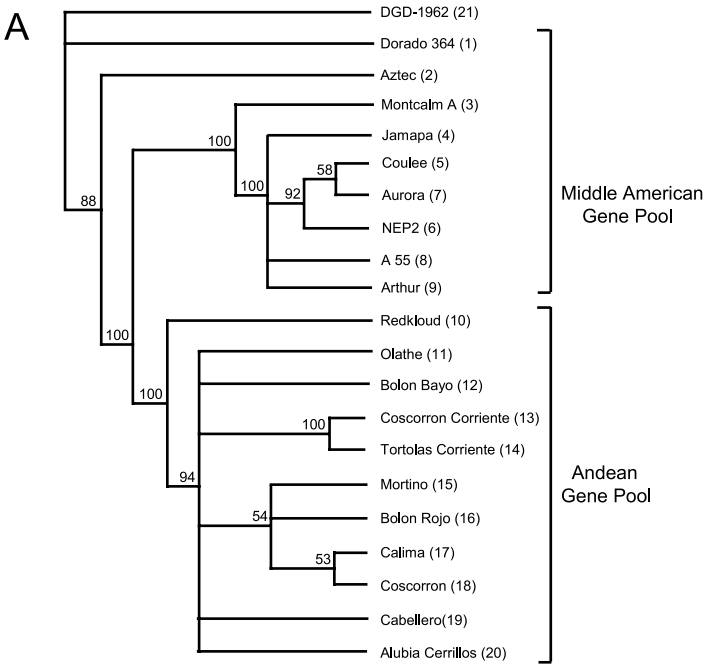
From the sequencing results, the *Aurora* allele was less frequent than the *Jamapa* allele. Our marker screening with the *Aurora* specific primers only identified one additional genotype, great northern 'Beryl', with this allele. Pedigree analysis revealed that 'Aurora' was part of the 'Beryl' ancestry.

Finally, an allele-specific SNP primer set was designed to the unique *Dorado* allele. None of the tested genotypes contained this allele.

Common bean contains linked, duplicate copies of *DFR*

Two results suggested that DFR might be duplicated in the common bean genome. First, two genotypes were positive for two different alleles when the allele-specific primers were tested. One of these, breeding line OT9630-17, contains both 'Sierra' and 'Bill Z' in its pedigree. Marker analysis demonstrated that 'Sierra' was homozygous for the *Durango* allele, and sequence analysis determined that 'Bill Z' was homozygous for the *Andean* allele. OT9630-17 was positive for both of these marker alleles.

Secondly, multiple sequencing attempts of the dark red kidney bean 'Montcalm' resulted in sequence data that became unreadable after only several hundred nucleotides. To



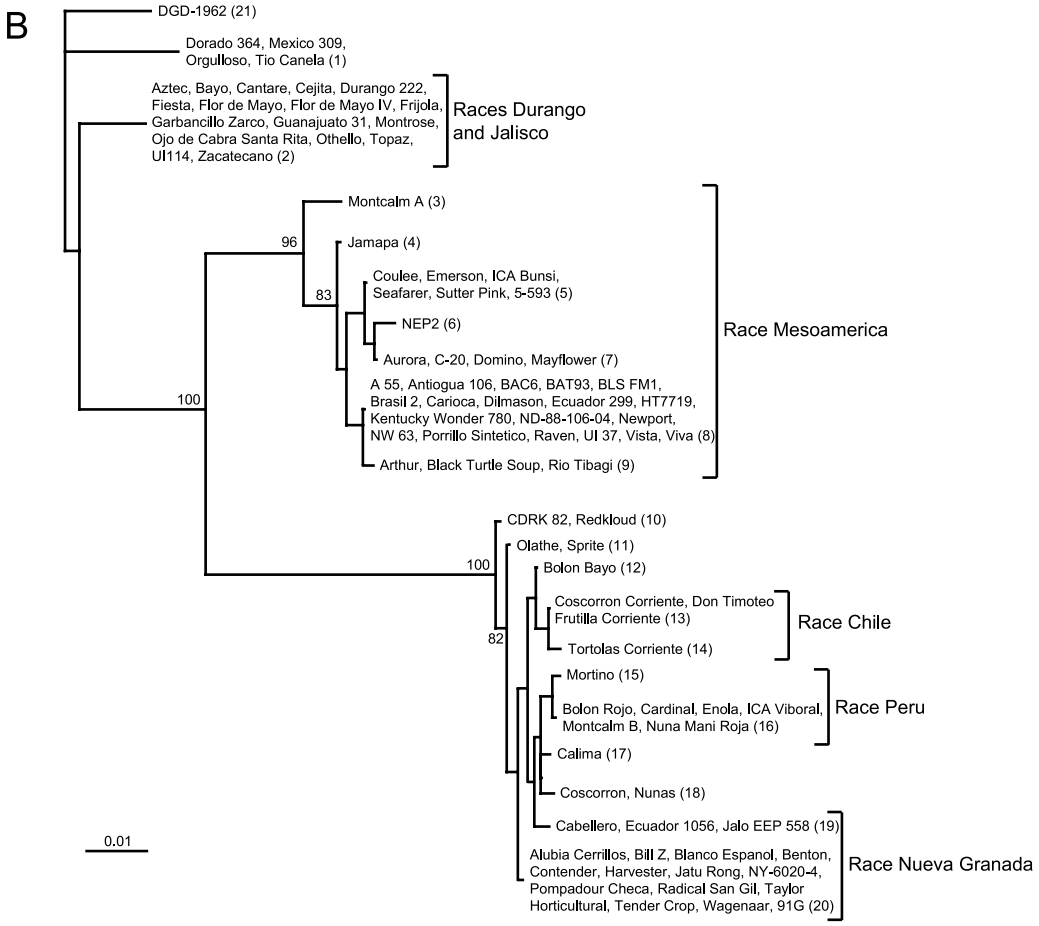
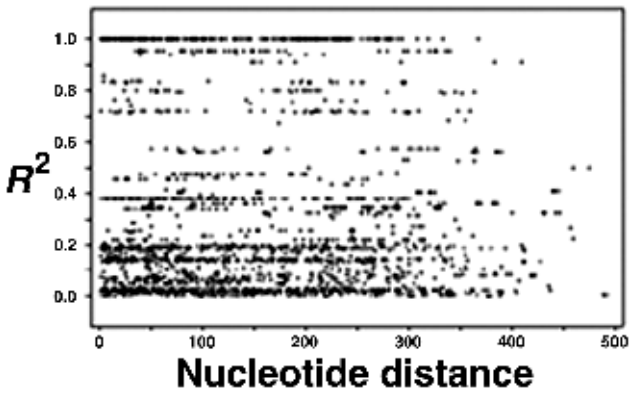


Fig. 3. A plot of R^2 (the linkage disequilibrium statistic of Hill and Robertson (1968)) versus nucleotide distance for intron 1 of dihydroflavonol 4-reductase of common bean.



obtain sequence data for this cultivar, we cloned the 'Montcalm' fragment and sequenced several clones. Two clone classes were obtained, and they were termed MontcalmA and MontcalmB. 'Montcalm' is a hybrid from the cross of 'Great Northern #1', a selection from the old landrace 'Common Great Northern', and 'Dark Red Kidney', an old cultivar from Michigan. Phenotypically, 'Common Great Northern' is classified as a member of the Middle American gene pool and 'Dark Red Kidney' is considered a member of the Andean gene pool. MontcalmA may descend from 'Common Great Northern', since it contains the deletion at nt 135–142 found in other Middle American genotypes. MontcalmB contains an insertion at this diagnostic site and may, therefore, have descended from 'Dark Red Kidney'. The marker data described above, and the sequencing data of Montcalm provide compelling evidence that DFR (as represented by intron 1) is a duplicate gene.

To determine the map location of the two loci, the *Durango* and *Dorado* allele-specific primer pairs were applied to a recombinant inbred mapping population (n = 79) developed from crossing 'Dorado' (*Dorado* allele) and 'Xan176' (*Durango*-331 allele). The population was screened, and the two loci mapped to the same linkage group with a multipoint distance of 15.9 cM (Fig. 4). Since one of the loci in this linkage group was the color gene *R* that is part of the complex *C* locus on linkage group B8, the two linked DFR loci map to that group.

Discussion

Applications of intron sequence data

The cloning, sequencing, and diversity analysis using DFR intron 1 has provided several important results. Although not entirely unexpected, we were able to show that the common bean intron is located at the same position in the gene as it is in several other plant species including both monocots and dicots. Using the same redundant primers, we successfully cloned the same intron from soybean and cowpea, and the intron location was also conserved in those species. This is important for the cloning of other introns because it suggests the intron–exon location data from a model species such as *Arabidopsis* can be used to select locations for primer design. If the primer location is not conserved among plant species, any given primer combination based on conserved amino acid sequences may inadvertently span an intron–exon border and amplification would fail. The full utility of our approach awaits the successful cloning of introns from other genes.

What remains to be determined by further research is whether other common bean introns contain the same level of diversity as DFR intron 1 and are, therefore, a rich source of data for diversity studies. Since this is the first extensive analysis of an intron in common bean, this question can only be addressed by comparing the diversity of this intron with that from other species. Diversity, as measured by π, was two to three times greater than the level found in chitinase genes in *Arabidopsis* (Kawabe et al. 1997; Kawabe and Miyashita 1999) and the two alcohol dehygrogenase genes of barley (Lin et al. 2001). The only plant gene with a greater reported level of intron diversity was *PgiC* of *Leavenworthia stylosa* (Filatov and Charlesworth 1999). This common bean intron also contained a greater number of polymorphic sites than found in any of the other species. Remarkably, this intron contained 69 polymorphic sites whereas Zhu et al. (2003) discovered only 96 SNPs among introns from 115 soybean genes. Although we are not comparing the exact same loci in each species, a testable hypothesis is that common bean introns are more diverse relative to other species.

Genetic diversity in common bean

The DFR intron 1 diversity data is in general agreement with previous studies of the genetic architecture of common bean germplasm. Both the MP and NJ trees and the data presented in the polymorphism table support the two gene pool concept (Gepts and Bliss 1986; Koenig and Gepts 1989; Gepts 1990; Becerra-Velásquez and Gepts 1994; Tohme et al. 1996). As with the previous molecular analysis studies, our data also show that the major divisions within the species are geographically based.

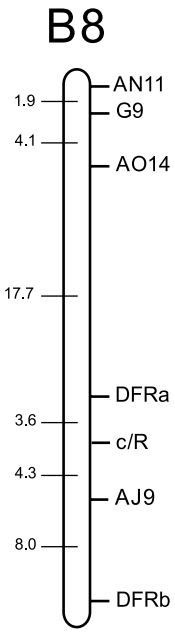
The wealth of polymorphisms at this locus revealed some similarities and differences with other diversity studies. Using RAPD analysis, Beebe et al. (2000) observed subsets within each Middle American race. As an example, race Mesoamerica was split into two subsets. M1 consisted of black Mesoamerican beans with upright indeterminate growth habit from Mexico, while subset M2 was comprised of genotypes from the southern range of the race with mostly non-black seed coat color and indeterminate prostrate growth habit. This analysis suggested that subtle differences within each Middle America race can be observed. Our data, which subdivided the race into six different haplotypes (5-9), is consistent with this observation. These haplotypes are closely related, as reflected by the short branch lengths on the NJ tree, and differ by only a few polymorphisms. Some of the haplotype groupings are reflective of breeding history. For example, 'NEP2' (haplotype 6) is a parent of 'C-20' (haplotype 7), and C-20 is one of the parents of two other haplotype 7 cultivars, 'Mayflower' and 'Domino'. These two haplotypes differ by two SNPs and one indel. Because we do not have sequence data for the other parents of these cultivars, we cannot determine if these polymorphisms are mutations that arose during the breeding process or were contributed by another parent in the pedigrees.

Table 5. DFR intron 1 allele-specific genotyping of common bean germplasm.

Genotype	Market class[a]	Allele				
		Andean	*Aurora*	*Dorado*	*Durango*	*Jamapa*
B190	Black	–	–	–	–	+
Benikintok	Coscorron, red	+	–	–	–	–
Citreon	Coscorron, yellow	+	–	–	–	–
Hutterite	Coscorron, yellow	+	–	–	–	–
Pariser Gelbe	Coscorron, yellow	+	–	–	–	–
Prim	Coscorron, yellow	+	–	–	–	–
VO 400	Coscorron, white	+	–	–	–	–
Beryl	Great northern	–	+	–	–	–
G97918	Great northern	–	–	–	+	–
G97943	Great northern	–	–	–	+	–
GN52	Great northern	–	–	–	+	–
GN1140	Great northern	–	–	–	–	+
GNB68	Great northern	–	–	–	+	–
Harris	Great northern	–	–	–	–	+
Matterhorn	Great northern	–	–	–	+	–
NE1-98-10	Great northern	–	–	–	–	+
Weihing	Great northern	–	–	–	–	+
UI59	Great northern	–	–	–	–	+
Akasanda	Kidney (red)	+	–	–	–	–
Amanda	Kidney (white)	+	–	–	–	–
Jacob's Cattle	Kidney (speckled)	+	–	–	–	–
L-98	Kidney (white)	+	–	–	–	–
Prakken 75	Kidney (white)	+	–	–	–	–
VO 687	Kidney (red)	+	–	–	–	–
CELRK	Kidney (red)	+	–	–	–	–
Early Gallatin	Snap (white)	+	–	–	–	–
Golden Gate Wax	Snap (white)	+	–	–	–	–
Isabella	Kidney (red)	+	–	–	–	–
Linden	Kidney (red)	+	–	–	–	–
Masterpiece	Kidney (buff)	+	–	–	–	–
Opal	Snap (gray)	+	–	–	–	–
Redlands Greenleaf B	Snap (red)	+	–	+	+	+
Redlands Greenleaf C	Snap (red)	+	–	–	–	–
Redlands Pioneer	Snap (brown)	+	–	–	–	–
Stringless Green Refugee	Snap (mottled)	+	–	–	–	–
Thuringia	Kidney (white)	+	–	–	–	–
Earli-Red	Medium red	–	–	–	–	+
Hidatsa	Medium red	–	–	–	–	+
Hueter	Medium red	+	–	+	+	+
Rufus	Medium red	–	–	–	–	+
UI37	Medium red	–	–	–	–	+
Dubble Witte	Medium white	–	–	–	–	+
Wax Digion	Medium tan	–	–	–	–	–
Alberta Pink	Pink	–	–	–	–	+
L94C356	Pink	–	–	–	–	+
UI537	Pink	–	–	–	–	+
BDM RMR 15	Pinto	–	–	–	+	–
BDM RMR 16	Pinto	–	–	–	+	–
BDM RMR 17	Pinto	–	–	–	+	–
Buster	Pinto	–	–	–	+	–
Maverick	Pinto	–	–	–	+	–
Mestizo	Pinto	+	–	–	–	–
OT9630–17	Pinto	+	–	–	+	–
Pinto 66	Pinto	–	–	–	+	–
Sierra	Pinto	–	–	–	+	–
UI129	Pinto	–	–	–	+	–

[a]Coscorron is used as a generic term to describe bean seeds with a large spherical shape. Kidney is used as a generic term to describe bean seeds of various sizes and colors that possess a kidney shape.

Fig. 4. The B8 linkage group showing the genetic location of the duplicate dihydroflavonol 4-reductase loci in common bean.



Similarly Beebe et al. (2000) described subsets of the 'Durango' and 'Jalisco' races based on geographic distribution, molecular marker patterns, seed type and growth habit. We did not observe that pattern for *DFR* intron 1. Rather, the genotypes representing these two races were all members of haplotype 2. The NJ tree also shows that this haplotype is quite different from the race Mesoamerica and Andean haplotypes. This haplotype was most closely related to haplotype 1, but still differed from it by 16 SNPs and two indels. It is interesting to note that both the MP and NJ methods clustered the race Mesoamerica genotypes with the Andean gene pool, before the race 'Durango' and 'Jalisco' haplotype was added to the tree. Both of these nodes were also supported with high bootstrap values.

The small red beans, such as 'Dorado', are generally considered members of race Mesoamerica, primarily because of their seed size (Beebe et al. 1995). Our data though clearly indicate that this market class, represented only by haplotype 1, is distinct from the other race Mesoamerica beans at DFR intron 1 by a deletion at nt 373–375 and several SNPs. This is best illustrated in the two gene trees. The NJ tree clearly shows the large distance between this haplotype and the race Mesoamerica clade (haplotypes 5–9). Although the placement of the small red beans in the NJ tree is not supported by a high bootstrap value, its placement was highly repeated among the most parsimonious MP trees. Collectively, these results demonstrate that for *DFR* intron 1, the small red class of bean is genetically unique.

Variation among the Andean genotypes at *DFR* intron 1 was greatly reduced relative to the Middle American genotypes. Even with this limited variation, the gene trees show the genotypes did cluster in a pattern similar to their race designation. Using molecular markers, Beebe et al. (2001) also observed reduced variation among Andean genotypes, but did not observe a race substructure.

It is presumed that the two gene pools of common bean have a common ancestor. It has been suggested that wild bean populations from northwest South America, which is geographically intermediate between the Middle America and the Andean region, may have been the source material for the two gene pools (Gepts 1998). Isozyme studies provided the first data to support this hypothesis (Koenig and Gepts 1989; Debouck et al. 1993). This was in turn supported by the discovery of wild beans with a phaseolin storage protein type intermediate to that found in the two gene pools (Kami et al. 1995). To investigate this concept further with the *DFR* locus, we included the wild bean DGD-1962 in our analysis as a representative genotype from the phaseolin study. The 40 amino acids of exon 2 of the wild bean were identical to those in all of the common bean genotypes studied. Yet analysis of the intron revealed that DGD-1962 was unique among this population. It contains the major deletion characteristic of the Middle American gene pool, but also exhibits five singleton SNPs and a unique single nt insertion. The unrooted phylogenetic trees (not shown) both reveal that DGD-1962 clusters at a common node with the 'Dorado' haplotype. The NJ tree shows long branch lengths separate these two haplotypes. The DFR sequence data, though, do not suggest DGD-1962 is ancestral between the two pools.

Our population was selected to include both cultivars and landraces because we were interested in determining the level of cultivar diversity relative to that represented by landraces. This would provide an indication of the extent of introgression of common diversity into the hybrid form of the species. As a group, the cultivars were as diverse as the landraces. This trend was also observed when the relationship was studied within each gene pool. This suggests that breeding efforts, at least for this genomic region, have successfully represented the variability in landraces in the cultivated genotypes. This can also been seen in the phylogenetic trees. Nearly all haplotypes that contain cultivars, also contain a landrace. The only major exceptions are haplotypes 5–7. It would be useful to study the pedigree of these genotypes in attempts to uncover a relevant landrace to add to our population for further analysis.

Evolution of variation in common bean

We are interested in assessing the contributions of mutation and recombination, the two forces that can generate variation, in the development of variation in common bean. Recombination requires outcrossing. Although common bean is a self-pollinated species, outcrossing has been documented. Although the degree of outcrossing is generally low (Triana et al. 1993; Ferreira et al. 2000), outcrossing rates of 66.8% have been reported (Wells et al. 1988). One place that outcrossing is documented in nature is in wild-weed-crop complexes. Recently, Beebe et al. (1997) described such complexes, and their genetic analyses conclusively showed that crossing occurs among the three states. Therefore, since mechanisms do exist for hybridization within common bean, we considered the role of recombination at the DFR locus.

We only focused on the landraces to address this issue because cultivars are the result of intentional hybridization (and thus recombination) and therefore are not informative from an evolutionary context. Six recombination events were

observed among the landraces. Yet when landraces within a gene pool were studied, recombination was not observed. This implies that variation within each gene pool is generated by mutation and not recombination. Relative to all of the landraces, the variation is reduced 25% among Middle American landraces, and eight-fold among Andean landraces. Therefore, the Middle American gene pool has accumulated more mutation events at this locus than has the Andean gene pool. It is also possible that the Andean gene pool contains genetic factors that eliminate newly generated variation.

Because intragenic recombination was observed among all landraces representing the two gene pools, a common genetic base must have existed in the history of common bean. We analyzed the recombination history of the landraces to determine which haplotypes may represent parents of these recombination events. The Hudson and Kaplan (1985) test was applied to subsets of all of the haplotypes to determine the minimal set required to explain specific recombination events. From that minimal subset, hypotheses were drawn regarding the evolutionary history of this region of the genome.

Of the six recombination events, four (nt 143, 162; nt 175, 196; nt 332, 339; and nt 339, 402) can be explained by hybridization among the same subset of four haplotypes. This subset must include the Middle American haplotypes one and two, and either haplotype eight or nine, and any one of the Andean haplotypes. Because recombination was not observed within either gene pool, the parents for these recombination events must include an Andean and Middle American haplotype. Under this scenario, haplotype one and the Andean genotype would donate the parental chromosomes, and the two other Middle American haplotypes represent the recombinant chromosomes. The other two recombination events (nt 60, 81 and nt 423, 483) could not be explained by a single subset of haplotypes, but all possible subsets require that the parental chromosomes be donated by one Andean and one Middle American haplotype. These results suggest that the ancestors of the Middle American and Andean gene pools hybridized at some point in the history of common bean.

Kami et al. (1995) described the wild bean DGD-1962 as a potential intermediate donor to the common bean germplasm. At DFR intron 1, this genotype could substitute only for the nt 143,162 and nt 175, 196 recombination events. This partially supports the hypothesis of Gepts (1998) that wild *P. vulgaris* genotypes were ancestors of the domesticated form of the species. Yet a more thorough analysis of additional wild bean germplasm is needed to determine the degree to which the variability represented in the landraces is representative of the species as a whole. Those analyses will allow us to determine if recombination among wild beans actually predated that observed within this set of landrances. Such experiments will additionally test the Gept (1998) hypothesis.

In general, we have demonstrated that common bean variation is the result of recombination between ancestors in a common population followed by mutation within each gene pool. In addition, it suggests that recombination occurred among individuals within a population that contained DFR intron 1 haplotypes similar to those in the current common bean gene pool. This is in contrast to the phaseolin locus where it is suggested that tandem duplications were introduced into both intron and exon regions prior to the occurrence of the two current gene pools (Kami et al. 1995). Therefore, in its history common bean variation was generated by a combination of mutations, recombination, and duplication events that occurred within and among the current gene pools as well in the ancestral gene pool.

The next goal is to address this concept in a broader context. First, it will be important to determine if this general pattern exists among other genes. These genes should represent not only metabolic genes such as *DFR*, but also regulatory genes. Furthermore, an investigation of not only low copy genes, but also gene families should be pursued. Finally, the relationship between variation in common bean and that found in other *Phaseolus* species should also be considered. This may allow us to look at molecular events associated with *P. vulgaris* speciation.

Acknowledgements

We thank Mark Basset, University of Florida; Dermot Coyne and Jim Steadman, University of Nebraska; Ken Grafton, North Dakota State University; and Jim Myers, Oregon State University, for providing seed of selected genotypes used in this study. We appreciate the unpublished DFR sequence data provided by Thomas Davis, University of New Hampshire.

References

Acosta-Gallegos, J.A., Ibarra-Perez, F.J., Rosales-Serna, R., Fernandez-Hernandez, P., Castillo-Rosales, B., and Kelly, J.D. 2001. Registration of 'Mestizo' pinto bean. Crop Sci. **41**: 1650–1651.

Becerra-Velásquez, V.L., and Gepts, P. 1994. RFLP diversity in common bean (*Phaseolus vulgaris* L.). Genome, **37**: 256–263.

Beebe, S.E., Ochoa, I., Skroch, P., Nienhuis, J., and Tivang, J. 1995. Genetic diversity among common bean breeding lines developed for Central America. Crop Sci. **35**: 1178–1183.

Beebe, S.E., Toro, O., Gonzalez, A.V., Chacon, M.I., and Debouck, D.G. 1997. Wild-weed-crop complexes of common bean (*Phaseolus vulgaris* L.) in the Andes of Peru and Colombia, and their implications for conservation and breeding. Genet. Resour. Crop Evol. **44**: 73–91.

Beebe, S.E., Skroch, P.W., Thome, J., Duque, M.C., Pedraza, F., and Nienhuis, J. 2000. Structure of genetic diversity among common bean landraces of middle American origin based on correspondence analysis of RAPD. Crop Sci. **40**: 264–273.

Beebe, S.E., Gaitan, R.E., Duque, M.C., and Tohme, J. 2001. Diversity and origin of Andean landraces of common bean. Crop Sci. **41**: 854–862.

Brady, L., Bassett, M.J., and McClean, P.E. 1998. Molecular markers associated with *T* and *Z*, two genes controlling partly colored seed coat patterns in common bean. Crop Sci. **38**: 1073–1075.

Corpet, F. 1988. Multiple sequence alignment with hierarchical clustering. Nucleic Acids Res. **16**: 10 881 – 10 890

Coyne, D.P., Nuland, D.S., Lindgren, D.T., Steadman, J.R., Smith, D.W., Gonzales, J., Schild, J., Reiser, J., Sutton, L., Carlson, C., Stavely, J.R., and Miklas, P. 2000. 'Weihing' great northern disease-resistant dry bean. HortScience, **35**: 310–312.

Debouck, D.G., Toro, O., Paredes, O.M., Johnson, W.C., and Gepts, P. 1993. Genetic diversity and ecological distribution of

Phaseolus vulgaris in northwestern South America. Econ. Bot. **47**: 408–423.

Doyle, J.J., and Doyle, J.L. 1990. Isolation of plant DNA from fresh tissue. Focus, **12**: 13–15.

Drenkard, E., Richter, B.G., Rozen, S., Stutius, L.M., Angell, N.A., Mindrinos, M., Cho, R.J., Oefner, P.J., Davis, R.W., and Ausubel, F.M. 2000. A simple procedure for the analysis of single nucleotide polymorphisms facilitates map-based cloning in Arabidopsis. Plant Physiol. **124**: 1483–1492.

Ferreira, J.J., Alvarez, E., Fueyo, M.A., Roca, A., and Giraldez, R. 2000. Determination of outcrossing rates of *Phaseolus vulgaris* L. using seed protein markers. Euphytica, **113**: 259–263.

Filatov, D. A., and Charlesworth, D. 1999. DNA polymorphism, haplotype structure and balancing selection in the Leavenworthia PgiC locus. Genetics, **153**: 1423–1434.

Fu, Y.-X., and Li, W.-H. 1993. Statistical tests for neutrality of mutations. Genetics, **133**: 693–709.

Gepts, P. 1990. Biochemical evidence bearing on the domestication of *Phaseolus* (Fabaceae) beans. Econ. Bot. **44**: 22–38.

Gepts, P. 1998. What can molecular markers tell us about the process of domestication in common bean? *In* The origins of agriculture and crop domestication. *Edited by* D.B. Damania, J. Valkoun, G. Willcox, and C.O. Qualset. ICARDA, Aleppo, Syria. pp. 198–209.

Gepts, P., and Bliss, F.A. 1986. Phaseolin variability among wild and cultivated common beans (*Phaseolus vulgaris*) from Colombia. Econ. Bot. **40**: 469–478.

Gepts, P., and Debouck, D. 1991. Origin, domestication, and evolution of common bean (*Phaseolus vulgaris* L.). *In* Common beans: research for crop improvement. *Edited by* A. van Schoonhoven and O. Voyest. CAB International, Wallingford, U.K. and CIAT, Cali, Colombia. pp. 7–53.

Hartl, D.L., and A.G. Clark. 1997. Principles of Population Genetics. 3rd ed. Sinauer Associates, Sunderland, Mass.

Hill, W.G., and Robertson, A. 1968. Linkage disequilibrium in finite populations. Theor. Appl. Genet. **38**: 226–231.

Hudson, R.R., and Kaplan, N.L. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics, **111**: 147–164.

Kami, J., Becerra-Velasquez, V., Debouck, D.G., and Gepts, P. 1995. Identification of presumed ancestral DNA sequences of phaseolin *Phaseolus vulgaris*. Proc. Natl. Acad. Sci. U.S.A. **92**: 1101–1104.

Kawabe, A., Innan, H., Terauchi, R., and Miyashita, N.T. 1997. Nucleotide polymorphism in the acidic chitinase locus (*ChiA*) region of the wild plant *Arabidopsis thaliana*. Mol. Biol. Evol. **14**: 1303–1315.

Kawabe, A., and Miyashita, N.T. 1999. DNA variation in the basic chitinase locus (*ChiB*) region of the wild plant *Arabidopsis thaliana*. Genetics, **153**: 1445–1453.

Kelly, J.K. 1997. A test for neutrality based on interlocus associations. Genetics, **146**: 1197–1206.

Kelly, J.D., and Miklas P.N. 1999. Marker-assisted selection. *In* Common bean improvement in the twenty-first century. *Edited by* S.P. Singh. Kluwer Academic Publishers, Dordrecht, The Netherlands. pp. 93–123.

Kelly, J.D., Hosfield, G.L., Varner, G.V., Ubersax, M.A., and Taylor, J. 1999. Registration of 'Matterhorn' great northern bean. Crop Sci. **39**: 589–590.

Kimura, M. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge, Mass.

Koenig, R., and Gepts, P. 1989. Allozyme diversity in wild *Phaseolus vulgaris*: further evidence for two major centers of diversity. Theor. Appl. Genet. **78**: 809–817.

Lander, E.S., Green, P., Abrahamson, J., Barlow, A., Daly, M.J., Lincoln, S.E., and Newburg, L. 1987. MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics, **1**: 174–181.

Lewontin, R.C., and Kojima, K. 1960. The evolutionary dynamics of complex polymorphisms. Evolution, **14**: 458–472.

Lin, J.-Z., Brown, A.H.D., and Cleeg, M.T. 2001. Heterogenous geographic patterns of nucleotide sequence diversity between two alcohol dehydrogenase genes in wild barley (*Hordeum vulgare* subspecies *spontaneum*). Proc. Natl. Acad. Sci. U.S.A. **98**: 531–536.

McClean, P.E., Myers, J.R., and Hammond, J.J. 1993. Coefficient of parentage and cluster analysis of North American dry bean cultivars. Crop Sci. **33**: 190–197.

Nei, M. 1987. Molecular Evolutionary Genetics. Columbia University Press, New York, N.Y.

Rokas, A., and Holland, P.W.H. 2000. Rare genomic changes as a tool for phylogenetics. Trends Ecol. Evol. **15**: 454–459.

Rozas, J., and Rozas, R. 1999. DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics, **15**: 174–175.

Singh, S.P., Gepts P., and DeBouck, D.G. 1991. Races of common bean (*Phaseolus vulgaris*, Fabaceae). Econ. Bot. **45**: 379–386.

Staden, R. 1994. The Staden package. *In* Methods in molecular biology. *Edited by* A.M. Griffin and H.G. Griffin. Humana Press Inc., Totawa, N.J. 25:9–170.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphisms. Genetics, **123**: 585–595.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., and Higgins, D.G. 1997. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. **24**: 4876–4882.

Tohme, J, Gonzalez, D.O., Beebe, S., and Duque, M. 1996. AFLP analysis of gene pools of a wild bean core collection. Crop Sci. **36**: 1375–1384.

Triana, B.M., Iwanaga, M., Rubiano, H., and Andrade, M. 1993. A study of allogamy in wild *Phaseolus vulgaris*. Annu. Rept. Bean Improvement Coop. **36**: 20–21.

Watterson, G.A. 1975. On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7**: 256–276.

Wells, W.C., Isom, W.H., and Waines, J.G. 1988. Outcrossing rates in six common bean lines. Crop Sci. **28**: 177–178.

Weir, B.S. 1996. Genetic Data Analysis II. Sinauer Associates, Sunderland, Mass.

Zhu, Y.L., Song, Q.J., Hyten, D.L., van Tassell, C.P., Matukumalli, L.K., Grimm, D.R., Hyatt, S.M., Fickus, E.W., Young, N.D., and Cregan, P.B. 2003. Single-nucleotide polymorphisms in soybean. Genetics, **163**: 1123–1134.